

Unsupervised Learning

Nipun Batra

July 19, 2025

IIT Gandhinagar

The need for Unsupervised Learning

- Aids the search of patterns in data.

The need for Unsupervised Learning

- Aids the search of patterns in data.
- Find features for categorization.

The need for Unsupervised Learning

- Aids the search of patterns in data.
- Find features for categorization.
- Easier to collect unlabeled data.

The need for Unsupervised Learning

- Aids the search of patterns in data.
- Find features for categorization.
- Easier to collect unlabeled data.

The need for Unsupervised Learning

- Aids the search of patterns in data.
- Find features for categorization.
- Easier to collect unlabeled data.

Places where you will see unsupervised learning

- It can be used to segment the market based on customer preferences.

The need for Unsupervised Learning

- Aids the search of patterns in data.
- Find features for categorization.
- Easier to collect unlabeled data.

Places where you will see unsupervised learning

- It can be used to segment the market based on customer preferences.
- A data science team reduces the number of dimensions in a large data set to simplify modeling and reduce file size.

Clustering

AIM: To find groups/subgroups in a data set.

AIM: To find groups/subgroups in a data set.

REQUIREMENTS: A predefined notion of similarity/dissimilarity.

AIM: To find groups/subgroups in a data set.

REQUIREMENTS: A predefined notion of similarity/dissimilarity.

Examples:

Market Segmentation: Customers with similar preferences in the same groups. This would aid in targeted marketing.



gt_iris.png

K-Means Clustering

K-Means Clustering

K-Means Clustering

- N points in a R^d space.

K-Means Clustering

- N points in a R^d space.
- C_i : set of points in the i^{th} cluster.

K-Means Clustering

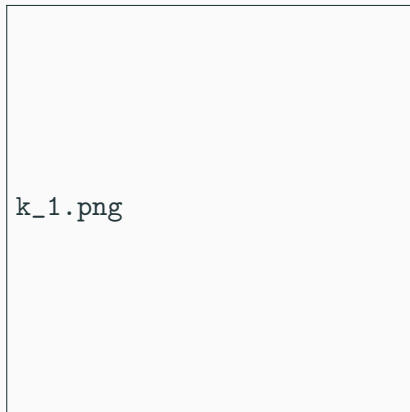
- N points in a R^d space.
- C_i : set of points in the i^{th} cluster.
- $C_1 \cup C_2 \cup \dots C_k = \{1, \dots, n\}$

K-Means Clustering

- N points in a R^d space.
- C_i : set of points in the i^{th} cluster.
- $C_1 \cup C_2 \cup \dots C_k = \{1, \dots, n\}$
- $C_i \cap C_j = \{\phi\}$ for $i \neq j$

K-Means Clustering

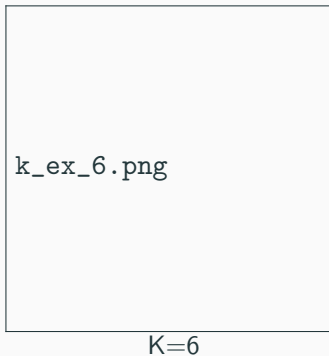
- N points in a R^d space.
- C_i : set of points in the i^{th} cluster.
- $C_1 \cup C_2 \cup \dots \cup C_k = \{1, \dots, n\}$
- $C_i \cap C_j = \{\phi\}$ for $i \neq j$



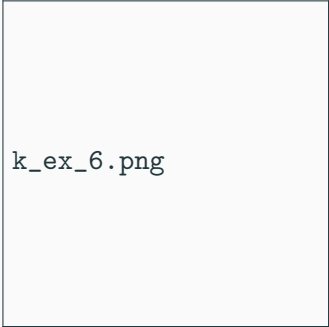
Dataset with 5 clusters

K-Means Clustering

K-Means Clustering



K-Means Clustering



k_ex_6.png

K=6



k_ex_5.png

K=5

K-Means Clustering



k_ex_6.png

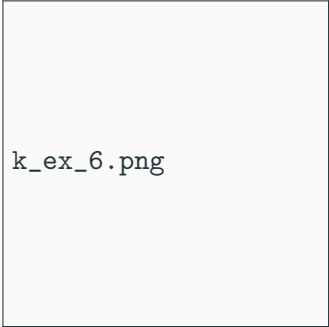
K=6



k_ex_5.png

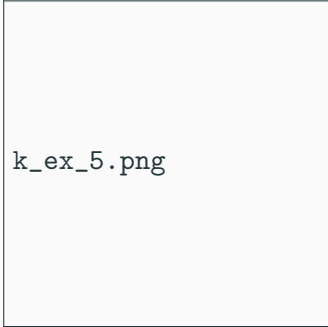
K=5

K-Means Clustering



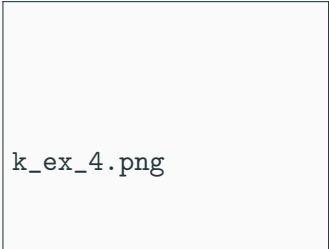
k_ex_6.png

K=6



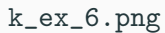
k_ex_5.png

K=5



k_ex_4.png

K-Means Clustering

A placeholder for a visualization of K-Means clustering with K=6. The image content is not visible.

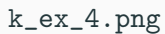
k_ex_6.png

K=6

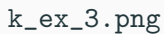
A placeholder for a visualization of K-Means clustering with K=5. The image content is not visible.

k_ex_5.png

K=5

A placeholder for a visualization of K-Means clustering with K=6. The image content is not visible.

k_ex_4.png

A placeholder for a visualization of K-Means clustering with K=5. The image content is not visible.

k_ex_3.png

K-Means Intuition

- Good Clustering: Within the cluster the variation (WCV) is small.

K-Means Intuition

- Good Clustering: Within the cluster the variation (WCV) is small.
- Objective:

$$\min_{C_1, \dots, C_k} \left(\sum_{i=1}^k WCV(C_i) \right)$$

K-Means Intuition

- Good Clustering: Within the cluster the variation (WCV) is small.
- Objective:

$$\min_{C_1, \dots, C_k} \left(\sum_{i=1}^k WCV(C_i) \right)$$

Minimize the WCV as much as possible

K-Means Intuition

Objective:

$$\min_{C_1, \dots, C_k} \left(\sum_{i=1}^k WCV(C_i) \right)$$

K-Means Intuition

Objective:

$$\min_{C_1, \dots, C_k} \left(\sum_{i=1}^k WCV(C_i) \right)$$

$$WCV(C_i) = \frac{1}{|C_i|} \text{ (Distance between all points)}$$

$$WCV(C_i) = \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} \|x_a - x_b\|_2^2$$

where $|C_i|$ is the number of points in C_i

K-Means Algorithm

1. Randomly assign a cluster number i to every point
(where $i \in \{1, \dots, n\}$)

K-Means Algorithm

1. Randomly assign a cluster number i to every point
(where $i \in \{1, \dots, n\}$)
 - 2.1 For each cluster C_i compute the centroid (mean of all points in C_i over d dimensions)

K-Means Algorithm

1. Randomly assign a cluster number i to every point
(where $i \in \{1, \dots, n\}$)
 - 2.1 For each cluster C_i compute the centroid (mean of all points in C_i over d dimensions)
 - 2.2 Assign each observation to the cluster which is the closest.

K-Means Algorithm

1. Randomly assign a cluster number i to every point
(where $i \in \{1, \dots, n\}$)
2. Iterate until convergence:
 - 2.1 For each cluster C_i compute the centroid (mean of all points in C_i over d dimensions)
 - 2.2 Assign each observation to the cluster which is the closest.

Working of K-Means Algorithm

Why does K-Means work?

Let, $x_i \in R^d =$ Centroid for i^{th} cluster

$$= \frac{1}{|C_i|} \sum_{a \in C_i} x_a$$

Why does K-Means work?

Let, $x_i \in R^d =$ Centroid for i^{th} cluster

$$= \frac{1}{|C_i|} \sum_{a \in C_i} x_a$$

Then,

$$\begin{aligned} WCV(C_i) &= \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} \|x_a - x_b\|_2^2 \\ &= 2 \sum_{a \in C_i} \|x_a - x_i\|_2^2 \end{aligned}$$

Why does K-Means work?

Let, $x_i \in R^d = \text{Centroid for } i^{\text{th}} \text{ cluster}$

$$= \frac{1}{|C_i|} \sum_{a \in C_i} x_a$$

Then,

$$\begin{aligned} WCV(C_i) &= \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} \|x_a - x_b\|_2^2 \\ &= 2 \sum_{a \in C_i} \|x_a - x_i\|_2^2 \end{aligned}$$

This shows that K-Means gives the **local minima**.

Hierarchal Clustering

Hierarchical Clustering

Gives a clustering of all the clusters

Hierarchical Clustering

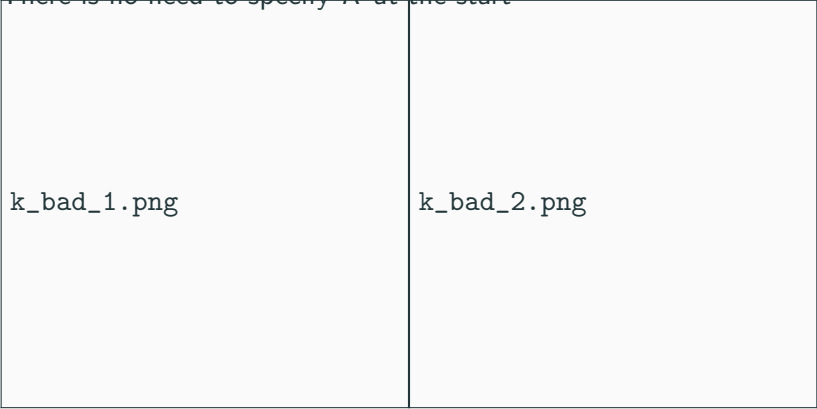
Gives a clustering of all the clusters

There is no need to specify K at the start

Hierarchical Clustering

Gives a clustering of all the clusters

There is no need to specify K at the start



k_bad_1.png

k_bad_2.png

Examples where K-Means fails

Algorithm for Hierarchical Clustering

1. Start with all points in a single cluster

Algorithm for Hierarchical Clustering

1. Start with all points in a single cluster

- 2.1 Identify the 2 closest points



h_e_1.png

Algorithm for Hierarchical Clustering

1. Start with all points in a single cluster

2.1 Identify the 2 closest points

2.2 Merge them

h_e_1.png

Algorithm for Hierarchical Clustering

1. Start with all points in a single cluster
2. Repeat until all points are in a single cluster
 - 2.1 Identify the 2 closest points
 - 2.2 Merge them

h_e_1.png

h_e_2.png

Joining Clusters/Linkages

Complete

Max inter-cluster
similarity

Complete

Max inter-cluster
similarity

Single

Min inter-cluster
similarity

Joining Clusters/Linkages

Complete

Max inter-cluster
similarity

Single

Min inter-cluster
similarity

Centroid

Dissimilarity between
cluster centroids

[Google Colab Link](#)